DOCUMENT RESUME

ED 232 594 HE 016 518

AUTHOR Cohen, Margaret E.
TITLE Use of Automatic Interaction Detector in Monitoring Faculty Salaries. AIR 1983 Annual Forum Paper.
PUB DATE May 83
NOTE 22p.; Paper presented at the Annual Forum of the Association for Institutional Research (23rd, Toronto, Ontario, May 23-26, 1983).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Faculty; Data Analysis; Employment Practices; *Evaluation Methods; Higher Education; Institutional Research; Intellectual Disciplines; Predictor Variables; *Salary Wage Differentials; *Teacher Characteristics; *Teacher Salaries
IDENTIFIERS *AIR Forum

ABSTRACT

A university's use of the Automatic Interaction Detector (AID) to monitor faculty salary data is described. The first step consists of examining a tree diagram and summary table produced by AID. The tree is used to identify the characteristics of faculty at different salary levels. The table is used to determine the explanatory power of the selected predictors and to determine the total amount of variation explained by the analysis. The second step consists of identifying the faculty whose salaries differ significantly from the means of their final groups and then reviewing their individual situations. The final step involves the use of a pairing method in situations where apparent inequity has been flagged. The main advantage of AID is its ability to detect interactions among variables. The application of the AID program was faculty data for 1979-1980 through 1981-1982 indicated that field was the most explanatory predictor of salary. Since most of the female faculty were in the fields having the lower salaries, the all-university average salary for females was lower than that of the male faculty. However, within fields there were still differences between salary averages. Race, ethnic origin, or sex were not selected as significant predictors of salary. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED232594

USE OF AUTOMATIC INTERACTION DETECTOR IN MONITORING FACULTY SALARIES

Margaret E. Cohen

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

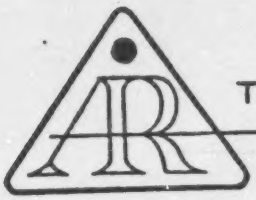
AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

HE016518

Abstract

Institutions must be able to provide information on the equity of faculty salaries. If the characteristics of the faculty at different salary levels can be identified, it is possible to determine whether or not a pattern of bias exists in the overall salary structure. The Automatic Interaction Detector (AID) segments a total group into mutually exclusive subgroups by the variables which explain the largest significant differences in a stated dependent variable. By using salary as the dependent variable and various attributes of the faculty as the independent variables, the AID can identify characteristics of the faculty who are paid at different levels. Since the mean of each final group is the predicted value of the observations in the group and since the AID supplies the normalized residuals for each observation in the final groups, individuals whose salaries are high or low for their groups can be identified for review. This paper describes one university's use of the AID to determine whether or not faculty salaries are equitable.



THE ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper was presented at the Twenty-Third Annual Forum of the Association for Institutional Research held at the Sheraton Centre in Toronto, Ontario, May 23-26, 1983. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum papers.

D. R. Coleman, Chairman
Forum Publication
Advisory Committee

Use of the Automatic Interaction Dectector in Monitoring Faculty Salaries

Monitoring equity in faculty salaries is an important aspect of an Equal Opportunity Program. Monitoring is required by federal regulation, but perhaps more important is the need to communicate effectively with faculty members about the differences in salary averages. At this university salary averages by school, rank, and sex get wide distribution. These data demonstrate that differences do exist between average salaries for male and female faculty members. They give rise to the question of why, if salary equity exists, these differences should remain.

The literature contains many studies addressing the issue of salary equity. In 1977 the American Association of University Professors published the Higher Education Salary Evaluation Kit in which multiple regression is recommended as the method for flagging women and minority persons whose salaries appear inequitable (Scott, 1977). Morse (1979) demonstrates that comparisons of mean salaries by rank and sex can produce misleading results and then recommends the use of multiple regression models as an alternative. The legal justification for using multiple regression is explored by Greenfield (1977). Today the prevalent method of monitoring faculty salary equity employs multiple regression models.

However, the use of multiple regression models has some limitations. It entails making a priori specifications of joint effects (interactions) among the variables. Ramsay (1979) points out that often little attention is paid to the development of a correct model. An incorrectly specified model which

ignores the effects of interactions can lead to erroneous conclusions.

In contrast, the Automatic Interaction Detector (AID) only requires minimal model specification (the specification of variables to be included). It is a process which reveals the relationships among the data; it identifies interactions which exist. This paper describes one university's incorporation of the AID into its affirmative action program. The process is twofold. The first step is the application of the AID program to faculty data in order to identify the characteristics of faculty at different salary levels. The second step is the isolation, for further review, of the individuals whose salaries differ significantly from the means of their groups.

The Automatic Interaction Detector

Procedure

The AID algorithm, developed by Morgan and Sonquist (1964), is a process which partitions data into a series of mutually exclusive subgroups by the independent variables (or predictors) which explain the most variation in a given dependent variable. In so doing it identifies the variables that affect the formation of the subgroups and reveals the existence of interactions among them. The AID is an iterative procedure which begins with the total sample input as the first group. This version[1] of the AID algorithm does the following:

1. Selects the unsplit group with the largest total sum of squares (TSS_i), starting with the total group (TSS_T).

2. For each predictor within that group, orders its categories by the

means of the dependent variable[2] and then considers all dichotomous groupings of the categories to select the one that produces the largest between sum of squares (BSS_i). For example, suppose highest degree is a predictor with categories 1-doctorate, 2-first professional, 3-masters, and 4-bachelors; and suppose further that the means of the dependent variable in these categories are ranked in like order. The groups tested will be: 1 versus 2, 3, and 4; 1 and 2 versus 3 and 4; and 1, 2, and 3 versus 4; the one with the largest BSS_i will be selected.

3. Selects the predictor from step 2 that maximizes the amount of explained variation (the predictor with the largest BSS_i/TSS_T).

4. Partitions the group into two subgroups by the following stopping rules: (a) a split eligibility criterion requiring that the ratio of the group's total sum of squares to the total sum of squares is greater than a user specified threshold value Q such that $TSS_i/TSS_T > Q$; (b) a split reducibility criterion requiring that the percentage of total variation explained by the split is larger than a user specified threshold value $P \geq Q$ such that $BSS_i/TSS_T > P$; (c) a test of the difference between the means of the two groups yields a t significant at a user specified level of significance; (d) the resultant subgroups are a minimum size; and (e) the total number of subgroups formed is less than 50.

Successive iterations of this procedure are applied to resultant subgroups until the criteria can no longer be met. At this point no further significant variation in the dependent variable can be explained by additional formation of minimum-size subgroupings of the specified variables.

This version of the AID produces a complete description of all the

attempted splits, a tree diagram showing the groups formed and identifying the variables by which the splits are made, and a table of summary statistics for each group formed. In addition, it optionally produces a complete description of the elements in the final groups including their normalized residuals.

Advantages and Disadvantages of AID

The major advantage of the AID is its ability to detect interactions among the variables. The structure of the data is evident from the pattern of the tree diagram (Sonquist and Morgan, 1964). If the tree is asymmetrical, then the data are not additive; interaction effects are present. If the tree is symmetrical, then the data are additive. A symmetrical pattern will not only show the same predictors used at parallel branches of the tree, but the explanatory power of each predictor will be similar. In either case, the variables with the most explanatory power are easily identified. If interaction effects are present, the AID provides information suggesting the forms these interactions should take. Sonquist (1970) has shown that the AID can identify interaction effects even in the presence of noise. Although it is possible to incorporate interaction terms in a multiple regression model, without prior knowledge of the forms these interactions should take the procedure can become unwieldly. Either all possible terms must be included or there is the possiblity of omitting significant ones.

Because the algorithm terminates in smaller mutually exclusive subgroups of the total group, each final group can be isolated for separate study. At that point effects which are unique to a specific group are identified.

One disadvantage of the AID is that when two predictors are highly intercorrelated, one of them may never be used (Sonquist, 1970). However,

with multiple regression both correlation coefficients for such inter-correlated variables might be close to zero "since, for each, a highly correlated variable is being 'held constant' in determining the net effects of the other" (p. 96).

Although the AID requires making minimal assumptions about the data, it does react to situations where a predictor is highly skewed or has a disproportionate number of classes (Sonquist, 1970). If a highly skewed predictor is not powerful enough to be selected early in the partitioning process, a less skewed predictor with similar power will be more likely to meet the minimum-size criterion for a later split. However, analysis of the residuals can uncover the effects of variables omitted because of skewness. When a predictor has an unusually large number of possible categories, spurious splits can occur. If this predictor is highly correlated with a powerful factor which has fewer classes, it can substitute for that factor. Careful regrouping of the categories can eliminate this possibility.

Another obvious drawback to the use of the AID is the necessity of specifying which variables to use as predictors of the dependent variable. This restriction is not unique to AID; it is inherent in other quantifiable techniques.

Data Sources

The source of data was the university's full-time faculty databank. Excluded from the analysis were administrators above the rank of assistant dean and research faculty.

The determination of which variables to use as predictors of salary was

made by the University's Commission on Equal Opportunity. Both rank and years in rank were excluded because these factors are subject to similar decision-making processes as those that determine salary (Scott, 1977). Thus their use might mask the discrimination that this analysis was attempting to reveal. The predictors used were: race/ethnic origin, sex, age, time at the university, highest degree, years since highest degree, field[3] of employment, and tenure track. All these factors except race/ethnic origin and sex were judged to be non-discriminatory explanations of salary differences. Age and years since highest degree were recoded into five-year intervals represented by numbers (1,2,3,...). Time at the university was also recoded into five-year intervals except the first interval which was seven years to reflect the time required prior to the tenure decision. Salaries were all standardized to an academic year basis (nine months).

The criteria used for a split to be made were: (a) the ratio of the group's total sum of squares to the total sum of squares must be greater than .5%; (b) the total variation explained by the split must be greater than .5%; (c) the significance level of the test of the difference between the means of the two groups was set at 5%; and (d) the resultant subgroups must have had at least 25 observations.

How were these criteria selected? (And would the use of another set change the outcome?)

Analysis

Figure 1 is the tree diagram produced from the 1981-82 faculty data. It is immediately apparent from the asymmetrical formation of the tree that interaction effects are present. For example, group 2 splits on age while group 3 splits on years since highest degree. Moreover, the right side of the

tree, the one with the higher average salaries, consists of three levels formed by field and age; the left side, the one with the lower average salaries, consists of five levels formed by years since highest degree, field, and age. By following any branch of the tree to its final group, the characteristics of the faculty in this group are identified. In this example, the lowest paid group with an average salary of $18,080 consists of the humanities faculty with less than ten years since receiving their highest degree. The highest paid group with an average salary of $52,100 consists of the law faculty and the medical faculty with MD degrees[4] who are aged 45 to 54 or aged 60 and over. Since the time variables were not constrained, the resultant subgroups were not forced to contain contiguous categories[5]. The discontinuous age groups formed from the 40 and over law and medicine faculty reflect the influence of market factors on their initial salaries.

Insert Figure 1 about here

The group summary table supplements the data available in the tree diagram. Table 1 is the summary table corresponding to the tree of figure 1. The groups are presented in order of formation giving summary statistics for each. The most informative statistic given is the column showing the cumulative amount of variation explained at each step of the partitioning process. The last value in this column is $\underline{R}^2$ for the segmentation analysis; in this case 66% of the total variation in salaries was explained by the variables used. The amount of variation explained by a specific variable at a given split can be determined. For example, the amount of variation explained

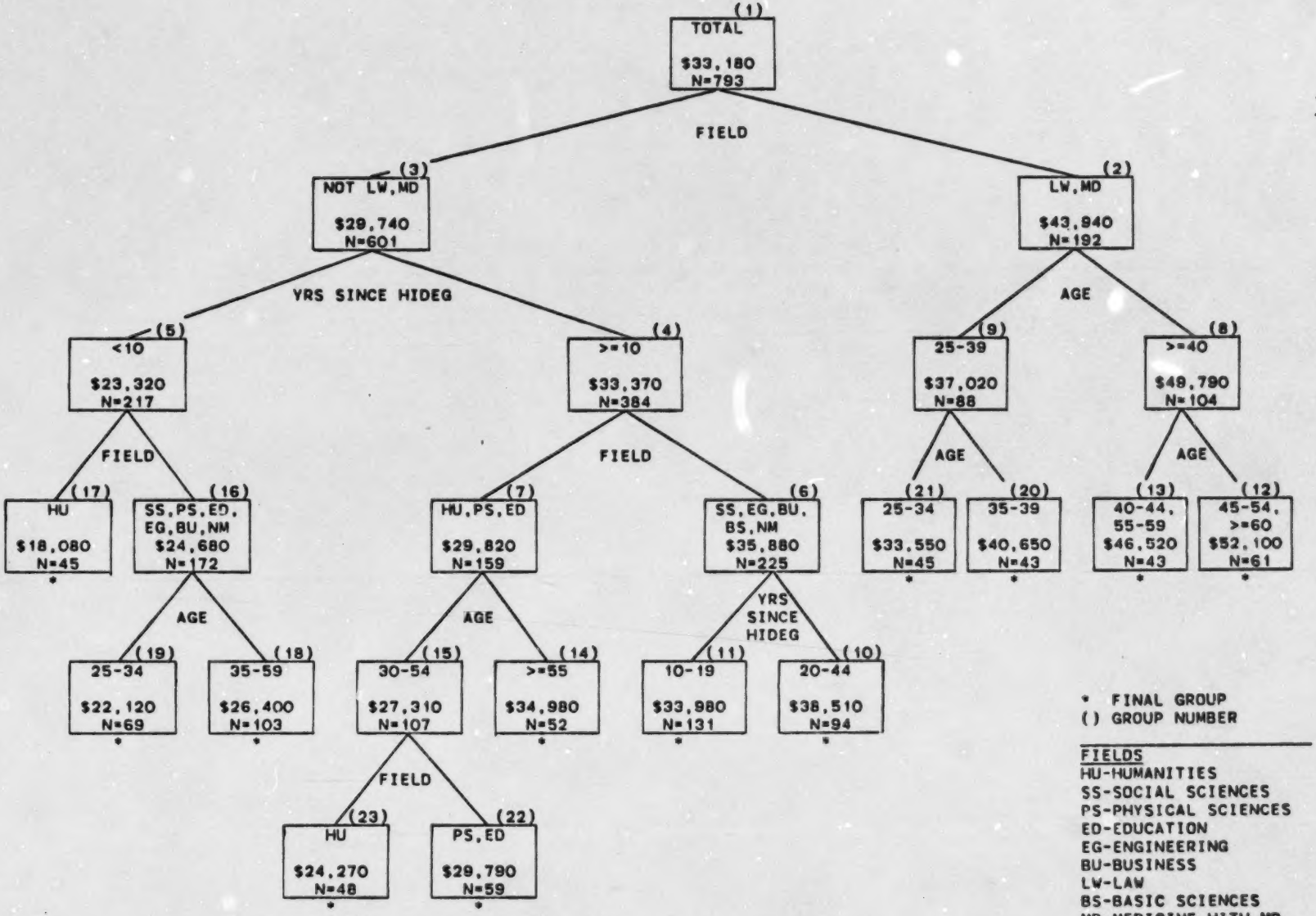


FIGURE 1. AID SEGMENTATION OF 1981-82 FACULTY SALARIES

by age at the group 2 split is 8% while that explained by years since highest degree at the group 3 split is 15%. These data can be transferred to the tree diagram to compare visually the explanatory power of the splits at parallel branches. The values can also be accumulated for each variable to determine its total explanatory power. In this example, field, years since highest degree, and age explain 38%, 16%, and 13%, respectively, of the total variation.

Insert Table 1 about here

Dichotomization terminates and final groups are formed when at least one of the splitting criteria is not met. If 49 groups have been formed, splitting terminates. However, only 23 groups are shown here. Since groups 13, 17, 20, 21, and 23 have fewer than 50 elements, they could not form additional groups with the minimum size of 25. From the total sum of sc res column on the summary table the split eligibility criterion can be calculated. In this example, each final group meets the requirement that its total sum of squares is at least .5% of the total sum of squares. The data necessary to examine the other criteria are not available from either the tree or the summary table.

The output from this AID program consisting of the program's analysis of each group provides sufficient detail to evaluate the other criteria. The detail includes identification of the most significant split that could be made for each predictor with the resultant subsample sizes, between sum of squares and t statistic. Groups 10, 14, and 22 would have split on highest

TABLE 1

GROUP SUMMARY TABLE OF 1981-82 FACULTY SALARY SEGMENTATION

GROUP	PARENT	N	MEAN	TSS	CUM R/TOT	SPLIT VARIABLE	CODES
1	-	793	3.318E+04	9.514E+10	0.00000	- ---	---
2	1	192	4.394E+04	2.050E+10	0.30838	7 FIELD	7 9
3	1	601	2.974E+04	4.530E+10	0.30838	7 FIELD	8 5 6 3 2 4 10 1
4	3	384	3.337E+04	2.39[illegible]E+10	0.45583	6 YRS SINCE HIDEG	6 7 8 4 5 9 3 2
5	3	217	2.332E+04	7.334E+09	0.45583	6 YRS SINCE HIDEG	1 0
6	4	225	3.588E+04	1.281E+10	0.49154	7 FIELD	8 5 6 2 10
7	4	159	2.982E+04	7.727E+09	0.49154	7 FIELD	3 4 1
8	2	104	4.979E+04	8.358E+09	0.57328	3 AGE	13 9 10 12 8 11
9	2	88	3.702E+04	4.370E+09	0.57328	3 AGE	7 6 5
10	6	94	3.851E+04	7.601E+09	0.58505	6 YRS SINCE HIDEG	6 4 7 5 8
11	6	131	3.398E+04	4.087E+09	0.58505	6 YRS SINCE HIDEG	3 2
12	8	81	5.210E+04	4.935E+09	0.59329	3 AGE	13 9 10 12
13	8	43	4.652E+04	2.639E+09	0.59329	3 AGE	8 11
14	7	52	3.498E+04	1.868E+09	0.61493	3 AGE	12 13 11
15	7	107	2.731E+04	3.801E+09	0.61493	3 AGE	10 9 8 7 6
16	5	172	2.468E+04	5.072E+09	0.63127	7 FIELD	5 6 10 4 8 2 3
17	5	45	1.808E+04	7.080E+08	0.63127	7 FIELD	1
18	16	103	2.640E+04	3.187E+09	0.63922	3 AGE	11 10 9 8 7
19	16	69	2.212E+04	1.129E+09	0.63922	3 AGE	6 5
20	9	43	4.065E+04	1.123E+09	0.65089	3 AGE	7
21	9	45	3.355E+04	2.136E+09	0.65089	3 AGE	6 5
22	15	59	2.979E+04	1.811E+09	0.65938	7 FIELD	3 4
23	15	48	2.427E+04	1.182E+09	0.65938	7 FIELD	1

degree, tenure track, and time at the university, respectively, had the resultant sample sizes been large enough. Groups 11, 12, 18, and 19 did not meet the split reducibility criterion: i.e., the amount of variation that would have been explained was less than .5%. If this criterion had been met, groups 12 and 18 would have split on sex indicating the need for further examination. Groups 11 and 19 would have split on field.

This detail of the analysis can also be used to determine the explanatory power of variables not used. Moreover, if additional examination of the data shows that another predictor was consistently a close candidate for selection, another run can be made excluding the selected variable from consideration.

The optional listing produced by this AID program identifies the elements in each of the final groups and shows the values of the predictors, the dependent variable, and the normalized residual for each one. These data can be used for the residual analysis recommended for skewed predictors. They can be examined to determine which elements differ significantly from the means of their groups. Since sample elements are identified by case number, it is necessary to have an ordered list of the total group to equate an individual with a case number. These individuals can then be flagged for further review.

Results

The AID program has been run on the faculty data for the years 1979-80, 1980-81, and 1981-82. To date the resultant tree diagrams have been asymmetrical indicating the presence of interaction effects. Different variables with different explanatory power formed parallel branches of the trees. Replication is a method of validating results from the AID (Sonquist, 1970).

As shown on Table 2, the percentage of total variation explained by these analyses has been about 67%. Although there was variability in the predictors selected from year to year, field consistently split the first group and had the most explanatory power. This is consistent with the hypothesis that market factors affect salaries in higher education (Tuckman, Gapinski, and Hagemann, 1977). Age and years since highest degree accounted for most of the remaining explained variation. In no case did the program select race/ethnic origin and sex as explanatory factors of salaries.

Insert Table 2 about here

Since these prohibitive factors were highly skewed variables, analysis of the residuals was indicated. The analysis was conducted on all final groups with particular attention paid to those in which the prohibitive factors might have been selected. The residual data were used to identify faculty members whose salaries differed from the means of their groups by two or more standard deviations[6]. Upon review of these individual situations it was found that qualitative factors which are difficult, if not impossible, to quantify generally explained the differences. These factors included: administrative assignment (e.g., department chairman), high or low performance rating, no terminal degree, and initial appointment at a high rank. They are consistent with Long's (1979) "nonquantifiable determiners of salary" (p. 145). As noted by Scott (1977), "the special quality of an individual's academic performance ...may not be measured...yet such special qualities may frequently determine the outcome of salary decisions" (p. 10).

Table 2

Explanatory Power of Variables Used to Predict Salaries

1979-80		1980-81		1981-82	
Predictor	Power	Predictor	Power	Predictor	Power
Field	.38	Field	.37	Field	.38
Age	.22	Age	.23	Yrs Since Hideg	.16
Yrs Since Hideg	.03	Yrs Since Hideg	.04	Age	.13
Hideg	.03	Hideg	.03		
		Yrs at Univ	.01		
R^2	.66		.68		.67

Conclusion

According to Sonquist (1970), "comparisons of the relative importance of predictors and the selection of those which most reduce predictive errors are necessary" (p. 9). The Automatic Interaction Detector does just that. Application of the AID did not select race/ethnic origin or sex as significant predictors of salary. Field was identified as the variable with the most explanatory power. Since most of the female faculty are in the fields having the lower salaries, the all-university average salary for females is lower than that of the male faculty. However, within fields there are still differences between salary averages. If the other selected predictors do not explain these differences and if examination of the final groups indicates the possibility that prohibicive factors affect salary averages, yet another method of analysis is applied to the selected individuals. In this instance, the matching system of comparing faculty salaries is employed.

In summary, our method of monitoring faculty salaries is a three-step process. The first consists of examining the tree diagram and summary table produced by the AID. The tree is used to determine the characteristics of the faculty at different salary levels. The table is used to determine the explanatory power of the selected predictors and to determine the total amount of variation explained by the analysis. The second step consists of identifying the faculty whose salaries differ significantly from the means of their final groups and then reviewing their individual situations. The final step consists of using the pairing method in situations where apparent inequity has been flagged.

The results of this technique are easily understood by faculty and

administrators. It has been adopted in lieu of multiple regression as a method of monitoring faculty salary equity. However, if multiple regression does in fact become mandatory, the results of the AID would be used in the first step, the model specification.

The AID is applicable to other areas of higher education. The Student Affairs Office on this campus utilizes results of the AID in roommate matching. Lay, Maguire, and Litton (1981) demonstrate its applicability in an Admissions office and suggest other campus offices that might benefit from its use. The results of the AID are useful in themselves or can be incorporated as the first step in other multivariate techniques.

Footnotes

[1]This version of the AID is a PL/1 program written by Dr. Philip W. Wirtz, Associate Director of the Social Research Group, George Washington University. The PL/1 source program is no longer available, but has been rewritten in SAS. Any inquiries about the program should be directed to Dr. Wirtz.

[2]Optionally, a predictor can be constrained. In this case the categories of the predictors are consecutively ordered and only contiguous categories are allowed in a subgroup.

[3]The categories for field of employment are: humanities, social sciences, physical sciences, educati , engineering, business, law, basic sciences, medicine with the MD degree, and medicine without the MD degree.

[4]The salaries used for the medicine faculty reflect only the guaranteed component of their salaries; any income from clinical practice is excluded.

[5]The program was rerun constraining the time variables. This did not lead to any other variable being selected in lieu of the time variable. In one case a final age split was not made resulting in a 1% reduction in $\underline{R}^2$.

[6]The analysis of the residuals is currently a manual process.

References

Greenfield, E. From Equal to Equivalent Pay: Salary Discrimination in Academia. Journal of Law and Education, January 1977, 6(1), 41-62.

Lay, R., Maguire, J., & Litten, L. Identifying Distinctive Groups in a College Applicant Pool. Paper presented at the annual meeting of the Association for Institutional Research, Minneapolis, Minnesota, May 1981.

Long, J. V. The Idiosyncratic Determiners of Salary Differences. In T. R. Pezzullo and B. E. Brittingham (Eds.), Salary Equity. Lexington, Massachusetts: Lexington Books, 1979.

Morse, P. K. Detection of Sex-Related Salary Discrimination: A Demonstration Using Constructed Data. In T. R. Pezzullo and B. E Brittingham (Eds.), Salary Equity. Lexington, Massachusetts: Lexington Books, 1979.

Ramsay, G. A. A Generalized Multiple Regression Model for Predicting College Faculty Salaries and Estimating Sex Bias. In T. R. Pezzullo and B. E Brittingham (Eds.), Salary Equity. Lexington, Massachusetts: Lexington Books, 1979.

Scott, E. L. Higher Education Salary Evaluation Kit. Washington, D.C.: American Association of University Professors, 1977.

Sonquist, J. A. Multivariate Model Building: The Validation of a Search Strategy. Ann Arbor, Michigan: Institute of Social Research, 1970.

Sonquist, J. A., & Morgan, J. N. The Determination of Interaction Effects. Monograph No. 35, Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan, 1964.

Tuckman, H. P., Gapinski, J. H., & Hagemann, R. P. Faculty Skills and the Salary Structure in Academe: A Market Perspective. American Economic Review, 1977, 67, 692-702.